Plant Cell Rep (2009) 28:649–661
DOI 10.1007/s00299-008-0661-3

GENETICS AND GENOMICS

Comparative sequence analysis for *Brassica oleracea* with similar sequences in *B. rapa* and *Arabidopsis thaliana*

Dan Qiu · Muqiang Gao · Genyi Li · Carlos Quiros

Received: 5 August 2008 / Revised: 14 October 2008 / Accepted: 9 December 2008 / Published online: 28 December 2008

Abstract We sequenced five BAC clones of *Brassica oleracea* doubled haploid 'Early Big' broccoli containing major genes in the aliphatic glucosinolate pathway, and comparatively analyzed them with similar sequences in *A. thaliana* and *B. rapa*. Additionally, we included in the analysis published sequences from three other *B. oleracea* BAC clones and a contig of this species corresponding to segments in *A. thaliana* chromosomes IV and V. A total of 2,946 kb of *B. oleracea,* 1,069 kb of *B. rapa* sequence and 2,607 kb of *A. thaliana* sequence were compared and analyzed. We found conserved collinearity for gene order and content restricted to specific chromosomal segments, but breaks in collinearity were frequent resulting in gene absence likely not due to gene loss but rearrangements. *B. oleracea* has the lowest gene density of the three species, followed by *B. rapa*. The genome expansion of the *Brassica* species, *B. oleracea* in particular, is due to larger introns and gene spacers resulting from frequent insertion of DNA transposons and retrotransposons. These findings are discussed in relation to the possible origin and evolution of the *Brassica* genomes.

Keywords Synteny · Gene mapping · Glucosinolates · Comparative genomics · Transposable elements

Communicated by R. Schmidt.

Electronic supplementary material The online version of this article (doi:10.1007/s00299-008-0661-3) contains supplementary material, which is available to authorized users.

D. Qiu · M. Gao · C. Quiros (✉)
Department of Plant Sciences, University of California, Davis, CA 95616, USA
e-mail: cfquiros@ucdavis.edu

Present Address:
M. Gao
Department of Agronomy, University of Kentucky, Lexington, KY 40546, USA

G. Li
Department of Plant Science, University of Manitoba, Winnipeg, MB R3T2N2, Canada

Introduction

Genome analyses in the model species *Arabidopsis thaliana,* is a useful tool for comparative genomic studies in the related *Brassica* genus which include important crop species. Comparative mapping between the genomes of crop plants and their respective model species is becoming a common approach for the identification of markers and candidate genes for mapping studies and to expedite positional gene cloning. Genome sequencing projects for *B. rapa* and *B. oleracea* are in the process, providing an opportunity to analyze and study the genome changes associated with the origin and evolution of these species in relation to *A. thaliana* (Ayele et al. 2005; Lim et al. 2006; Yang et al. 2006; Hong et al. 2006).

The genus *Brassica* includes three main cultivated species, *B. nigra*, $n = 8$; *B. oleracea*, $n = 9$; and *B. rapa*, $n = 10$, all of which function genetically as diploids. However, early evidence (Sikka 1940), indicated that these species are paleopolyploids, which is also the case for *A. thaliana* (Blanc et al. 2000). It is widely accepted that *Brassica* species and *A. thaliana* are diverged from a common ancestor of about 14.5–20.4 million years ago (Yang et al. 1999). The current thinking is that the common progenitor of the *Brassicaceae* had a basic genome of $n = 4$ chromosomes, which underwent a whole genome duplication 24–40 Mya producing a tetraploid species of

$2n = 4x = 16$ (Henry et al. 2006). The genome of this putative species was similar to the present genomes of *A. lyrata* and *Capsella rubella* ($2n = 16$, ~230 Mbp of DNA), from which presumable the genomes of the *Brassica* species derive (Schranz et al. 2006). *A. thaliana* evolved from this common ancestral species 4–5 Mya after it became diplodized and suffered general gene loss and chromosomal rearrangements including fusions or fissions resulting in the present genome of $n = 5$ and 157 Mb of DNA. (Johnston et al. 2005; Henry et al. 2006). Based on samplings of less than 2% of the genome, either by molecular marker map construction (Lagercrantz 1998; O'Neill and Bancroft 2000; Park et al. 2005; Parkin et al. 2005; Rana et al. 2004; Schmidt et al. 2003) and FISH (Lysak et al. 2005; Ziolkowski et al. 2006), it has been proposed that the *Brassica* diploid species are also evolved from the $2n = 4x = 16$ ancestral species after additional rounds of genome duplication, resulting in an hexaploid ancestor. This would explain in part the increase in DNA content from 230 Mbp to 529–696 Mbp (Johnston et al. 2005) reported for monogenomic cultivated *Brassica* species. However, it ignores the fact that ploidy changes are changes in chromosome number. The monogenomic *Brassica* species have already identical or similar chromosome numbers than those in the putative ancestral species ($n = 8$, 9 and 10). It also ignores the role of transposable elements which has been estimated to expand as much as 20% of the *B. oleracea* genome (Zhang and Wessler 2004). Little is known about the genomic structure of the *Brassica* diploid species. The two main cultivated species *B. oleracea* and *B. rapa* have diverged 7.3 ± 4 Mya (Wroblewski et al. 2000) and there is high synteny conservation for at least half of the chromosomes (Parkin et al. 2005). In the present study, we analyze similar sequences which harbor major glucosinolate genes in *B. oleracea*, *B. rapa* and *A. thaliana* in an effort to provide additional clues on the structure and evolution of *Brassica* genome.

Materials and methods

Brassica oleracea BAC sequencing

BAC clones B47M9, B67C16, B77C13, B59J16 and B16J1 originate from *B. oleracea* var. *italica* (broccoli) doubled haploid 'Early Big' library (Gao et al. 2004). These clones were selected because they harbor major genes in the aliphatic glucosinolate pathway. Clones B47M9, B67C16, B77C13 and B16J1 were outsourced for sequencing 454 Life Sciences (Bradford CT) using pyrosequencing. B59J16 was sequenced at the CA&ES Genomics Facility (CGF) in the UCD campus following traditional techniques (Gao et al. 2004). Gaps were filled by a combination of primer walking and shotgun sequencing of sub-clones at both the sides of gaps. The summary of assemble details is shown in Supplementary Table 1. Final error rate was estimated using CONSED which is less than 1 bp per 100 kb. These sequences were deposited in GenBank under the following accession numbers: B16J1 (EU579454), B67C16 (EU581950), B77C13 (EU579455), B47M9 (EU673963) and B59J16 (EU568372). For comparison, we added to the analysis, sequences of three other BACs of the same broccoli variety, B21H13 (Gao et al. 2004), B19N3 (Gao et al. 2005) and B21F5 (Gao et al. 2006). Additionally, we included *B. oleracea* contigs in the analysis sequenced by Town et al. (2006) by comparing them to their corresponding *B. rapa* sequences in the public domain: BAC clones KBrH015M19 (AC172876), KBrB077F22 (AC189466), KBrB063K02 (AC189420), KBrH1070K21 (DQ369749), KBrH093K03 (AC155347), KBrB021P11 (AC189261), KBrS005L11 (AC189638), KBrH077A05 (AC155343) and KBrB080C12 (AC189471). The same method was used to analyze the sequences of these BACs.

Sequence analysis and gene-prediction

The BAC sequence was analyzed for protein-coding genes with the following gene-prediction of *A. thaliana* software: GenScan (Burge and Karlin 1997) and TwinScan, by comparing conserved regions in the DNA of both species (Flicek et al. 2003). The sequence of BACs was aligned with its corresponding *A. thaliana* sequences with BLAST 2.2.9 (Altschul et al. 1997). The BAC sequence was also compared to *Arabidopsis, Brassica*, and *Oryza sativa* ESTs, cDNAs, and CDS using BLAST and FASTA with NCBI, AGI and TIGR database (www.tigr.org/tdb/e2k1/bog1/) to analyze gene conservation. The conserved regions were translated into protein and tBLASTn applied to the GenBank protein database to adjust exon–intron boundaries (05/01/2008). The transposable elements (TE) in the sequences were predicted and located with the program "RepeatMasker" (A.F.A. Smith and P. Green, unpublished data, http://www.repeatmasker.org/) and BLASTN and BLASTX searches to the GenBank database to find by comparison all types of reported transposable elements (05/01/2008). The 'bases masked' number is calculated from the total number of basepair masked sequences. The 'bases masked' include the retroelements, DNA Transposons, low-complexity DNA and simple repeats.

Results

Annotation of the five *B. oleracea* BACs harboring GSL genes

The five *B. oleracea* BACs B67C16, B47M9, B77C13, B59J16 and B16J1 sequenced were selected because they

harbor a major aliphatic glucosinolate (GSL) gene *BoGS-OH* (At2g25450), *BoCS-lyase* (At2g20610), *BoCYP79F1* (At1g16410), *BoGSL-PROb* (At1g18500) and *BoS-GT* (At1g24100), respectively. The annotation of the genes in these BACs is shown in Supplementary Tables 2, 3, 4, 5 and 6. The other three BAC clones previously sequenced and analyzed include B21H13 harboring *BoGSL-ALKa* and *BoGSL-ALKb*, B19N3 harboring *BoGSL-ELONG* and *BoGSL-ELOG-L* and B21F5 harboring *BoGSL-PRO* (Gao et al. 2004, 2005, 2006).

Characteristics of *B. oleracea* and *B. rapa* sequences in relation to *A. thaliana*

A total of 2,946 kb of *B. oleracea* sequence, including all eight broccoli BAC clones and the contigs published by Town et al. (2006), 1,069 kb of *B. rapa* sequence and 2,607 kb of corresponding *A. thaliana* sequence were compared and analyzed. Most of the comparative data could be generated between *B. oleracea* and *A. thaliana*, since corresponding *B. rapa* sequences were available only for half of the *B. oleracea* clones, B16J1, B21F5 B67C16. Table 1 shows a global summary of the features of these sequences for all three species. All the genes in the sequences of the three species were taken into account, however, only 48% of the *B. oleracea* and 71% of the *B. rapa* genes had counterparts in their corresponding *A. thaliana* sequences due to breaks in synteny. BLAST search in the *B. oleracea* and *B. rapa* databases for the absent genes indicated that at least 50% had homologs somewhere else in the genome (data not shown). The global gene comparison for the chromosomal segments compared revealed that on an average *B. rapa* genes were significantly longer (3,218 bp) than those in *B. oleracea* (2,721 bp) and the latter tended to be longer than in *A. thaliana* (2,310 bp), but the difference was not statistically significant. These differences were associated mostly to intron size, which was significantly larger in *B. rapa* compared to the other two species. Gene density was lowest for *B. oleracea* (one per 8.0 kb) followed by *B. rapa* (5.4 kb) and *A. thaliana* (one per 4.5 kb). This parameter was associated to gene spacer size which was larger in the *Brassica* species, *B. oleracea* in particular (Table 1). Taking into account only the genes conserving collinearity, gene size as well as exon size and number, were the same for all three species. However, intron size was different between *A. thaliana* and the *Brassica* species, but not between *B. oleracea* and *B. rapa*.

Annotation of the eight *B. oleracea* BAC sequence (745.8 kb) resulted in the construction of a total of 94 gene models (Table 2). These include the updated annotation of BACs B19N3, B21F5 and B21H13 previously reported (Gao et al. 2004, 2005, 2006). Considering all the eight clones, we could classify them by gene density. B21H13 has the highest density with 23 gene models in 101.5 kb and B47M9 has the lowest density with eight gene models in 104.6 kb (Table 2). A total of 89 gene models were annotated in 495.7 kb sequence of four *B. rapa* BAC clones that could be partially aligned to the four *B. oleracea* BAC clones listed at the beginning of this section (Table 2; Fig. 1). The *B. rapa* BACs had higher gene density than the *B. oleracea* BACs, in agreement with the global sequence comparison summarized in Table 1. In order to get a better picture of the alignment of the

Table 1 Features of eight *B. oleracea* and four *B. rapa* BAC clones

	B. oleracea BAC clone	*B. rapa* BAC clone	*Arabidopsis*
Sequence length	745.8 kb	495 kb	449 kb
G + C content overall (%)	36.35	36.81	35.8
Protein-coding DNA (%)	45.40	43.68	42.90
Non-coding region (%)	33	33.80	32.50
Total number of genes	94	89	213
Average gene size (bp)	2721*	3,218	2,310*
Average gene density (bp per gene)	8012*	5,436	4,450
Average number exons per gene	5.6	6.9	5.3
Average exon size (bp)	225**	272	258*
Average number introns per gene	4.6	5.9	4.3
Average intron size (bp)	206*	240**	153**
Average spacer size (bp)	6,368**	3,369	2,470
Average gene size (bp) for homologus gene	2,569	2,685	2,436
Average exon size (bp) for homologus gene	235	254	246
Average intron size (bp) for homologus gene	225	238	176*

* $P < 0.05$; ** $P < 0.01$

Table 2 Summary of features identified in eight *B. oleracea* BAC clones and four *B. rapa* BAC clones

Feature	*B. oleracea* BAC clone								*B. rapa* BAC clone				
	B16J1	B19N3	B21F5	B21H13	B47M9	B59J16	B67C16	B77C13	Total	KBrH015M19	KBr077F22	KBrB063K02	KBrH1070K21	Total
Sequence length (kb)	92.4	96.7	82.3	101.5	104.6	74.4	71.2	104.1	745.8	138	109.9	138.0	109.8	495
Gene models in *B. oleracea* or *B. rapa*	11	10	14	23	8	9	9	11	94	22	27	27	15	91
Gene density (kb/gene)	8.4	9.7	5.9	4.4	9.1	8.2	8.9	9.5	8.0	6.3	4.1	5.1	7.3	5.4
Sequence length in *A. thaliana* (kb)	32.4	84.2	45.0	118.1	35.1	57.6	43.5	32.8	448.6	137.2	317.7	321.5	84.9	861.2
Gene models in *A. thaliana*	16	19	14	37	12	17	13	14	139	39	71	54	23	187
Sequence ratio (*BrassicalArabidopsis*)	2.86	1.15	1.83	0.86	2.98	1.29	1.64	3.18	1.66	1.01	0.35	0.43	1.29	0.58
Number collinear genes	9	8	12	23	7	9	8	7	83	16	15	17	15	63
Collinear genes (%)	56	42	86	62	58	53	62	50	60	41	21	31	65	33
Gene fragments	4	4	4	0	1	2	0	3	19	1	0	0	1	2
Masked bp (%)	12.92	23.62	14.09	8.28	8.6	16.94	7.94	19.2	13	7.2	3.1	4.9	10.2	6.4
Predicted transposons	14	22	15	9	7	11	6	11	95	8	4	7	11	30
TE insert into collinearity region	2	9	5	3	2	3	1	4	29	1	2	0	4	7
Percentage	14	41	33	33	29	27	17	36	31	13	50	0	36	23

Fig. 1 The comparison map of *B. oleracea* and *Brassica rapa* BAC clones with *A. thaliana*. *Open right arrow* DNA transposons, *filled right arrow* retroelements, *open rectangle* gene fragments. *Vertical lines* indicates sequence contigs. The *triangle* by each gene model name indicates the coding strand of the gene

corresponding sequences of *A. thaliana* and *B. rapa*, the BAC clones of the latter species were aligned to the physical map of *A. thaliana* (Fig. 1).

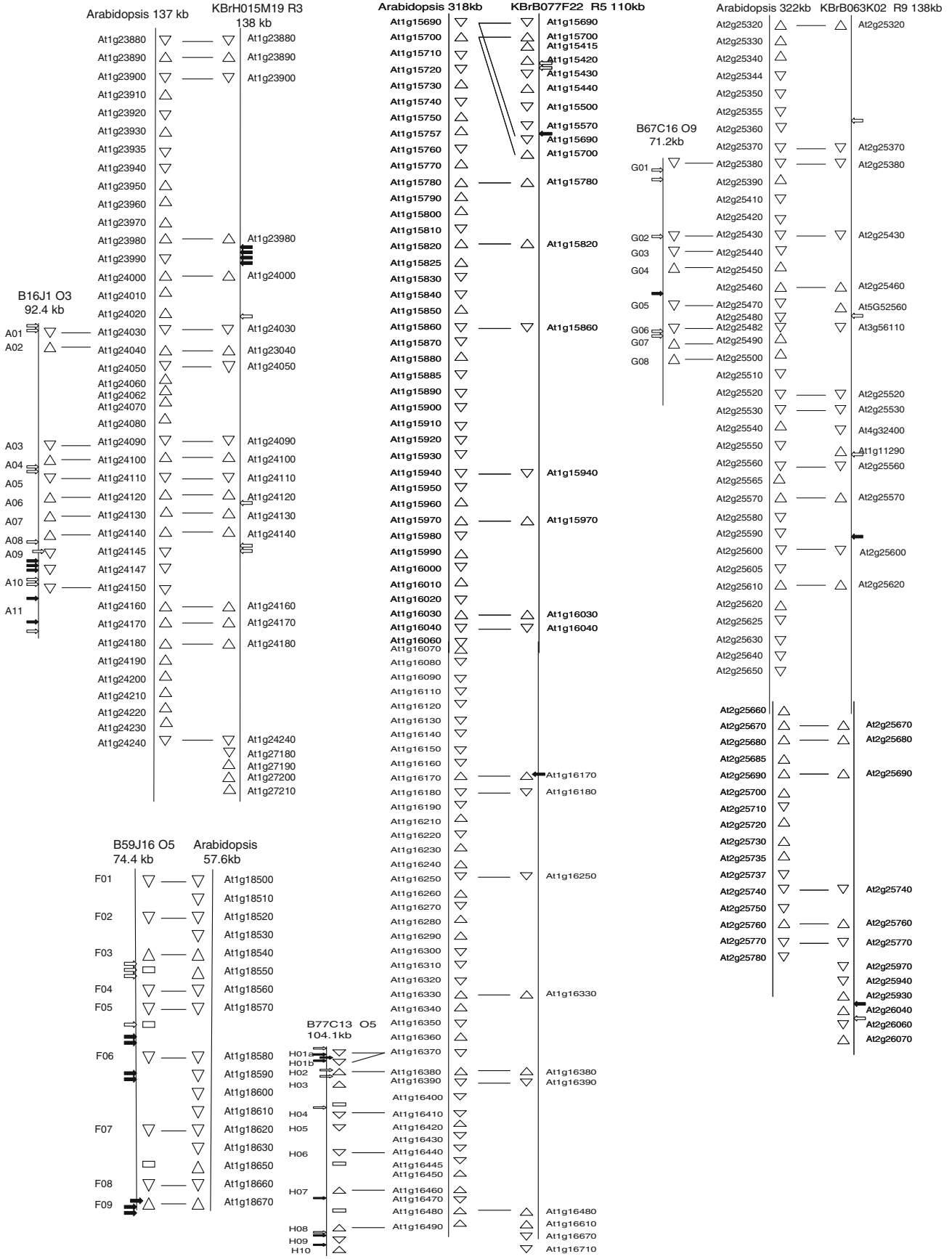
Annotation of *B. oleracea* contigs A B, C, D and G, covering a total of 1,518 Kb, resulted in the construction of 378 gene models (Table 3). Five *B. rapa* BAC clones could be partly aligned with these five *B. oleracea* contigs. A total of 60 gene models were annotated in 574 kb sequence of these corresponding *B. rapa* BAC clones (Table 3). Contrary to the trend found for the *B. oleracea* BAC clones, higher gene density were observed in the contigs for this species (4.0 kb) than in *A. thaliana* (4.6 kb) and *B. rapa* (5.6 kb).

All of these *Brassica* contigs have a high level of DNA sequence conservation with their counterparts in *A. thaliana*. One hundred and thirty nine and 187 *A. thaliana* gene models were identified in the corresponding region of eight *B. oleracea* and four *B. rapa* BAC clones, respectively (Fig. 1). Two hundred and thirteen and 136 *A. thaliana* gene models were identified in the corresponding region of five *B. oleracea* contigs and another five *B. rapa* BAC clones, respectively (Fig. 1).

DNA sequence conservation and collinearity between *Brassica* and *Arabidopsis*

In general, collinearity in the sense of finding corresponding genes in the same order and orientation was high among all the three species. In the eight *B. oleracea* BACs, 88% of the genes (83 of 94) conserved order with 139 *A. thaliana* genes in their corresponding regions (Table 2; Fig. 1). Sixty nine percent of *B. rapa* genes (63 of 91) conserved order with 187 *A. thaliana* genes in their corresponding regions (Table 2; Fig. 1). However, gene content was often different among all the three species in corresponding segments, due to frequent interspersed gene absence in the *Brassica* species in relation to *Arabidopsis* (Figs. 1, 2). Thus, when one considers in the comparison all the genes in the corresponding segments, the collinearity drops significantly to 60% between *B. oleracea* and *A. thaliana* and 33% between *B. rapa* and *A. thaliana* (Table 2). These values were higher for the five *B. oleracea* contigs, 70% (144/213) of the genes conserved collinearity to *A. thaliana* and 52% (73/136) for their corresponding *B. rapa* BAC sequences (Table 3; Fig. 2).

In a few instances genes present in chromosomal segments of *Brassica* species were absent in corresponding segments of *A. thaliana*. For example comparing


Arabidopsis 137 kb
KBrH015M19 R3
138 kb
B16J1 O3
92.4 kb
B59J16 O5
74.4 kb
Arabidopsis
57.6kb
B77C13 O5
104.1kb
Arabidopsis 318kb
KBrB077F22 R5 110kb
B67C16 O9
71.2kb
Arabidopsis 322kb
KBrB063K02 R9 138kb

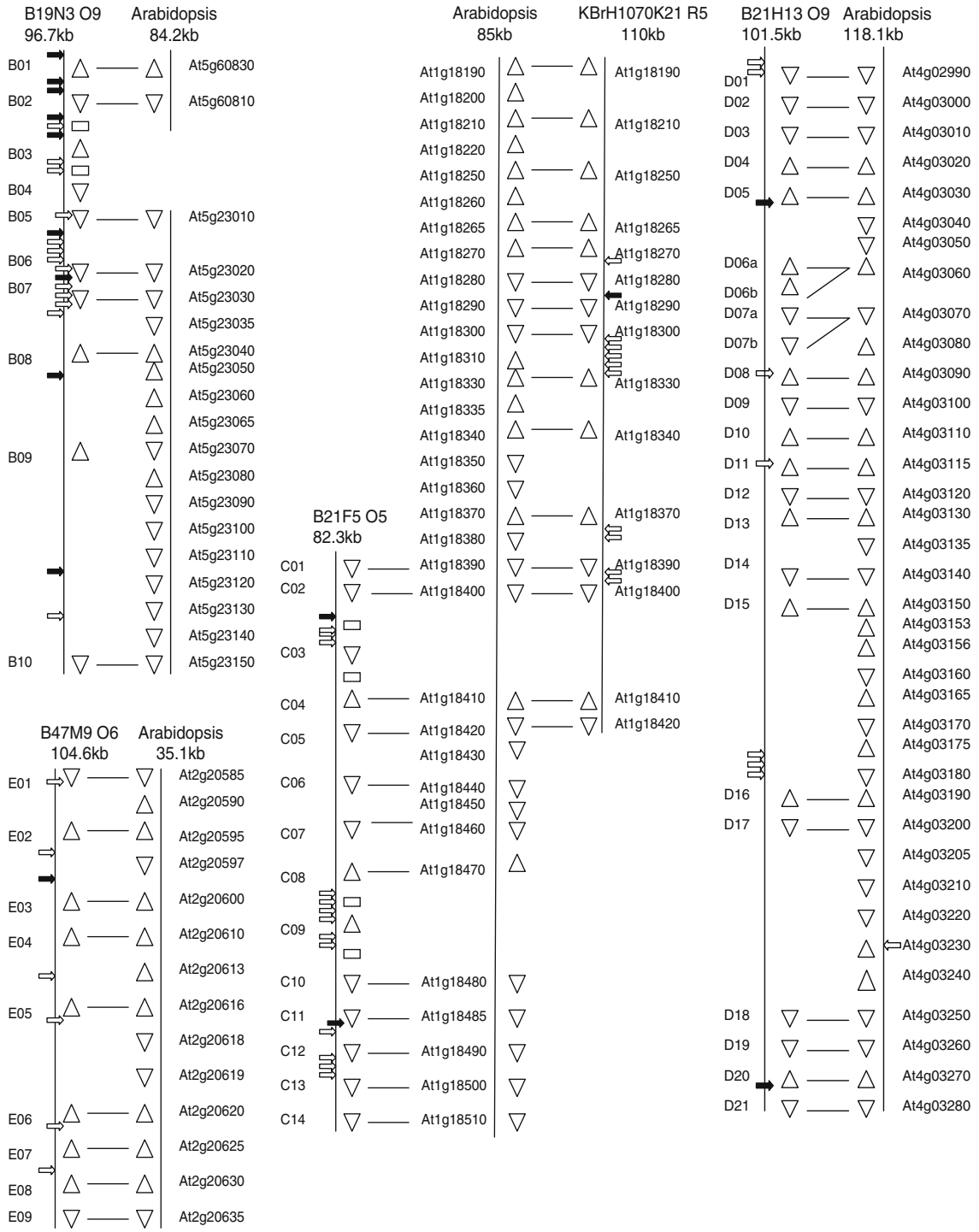


Fig. 1 continued

B. oleracea with *A. thaliana*, genes A09–A11 on B16J1, B09 on B19N3, C03 and C09 on B21F5, and H03, H05, H09 and H10 on B77C13 are absent in *A. thaliana*.

Often the genes showing the syntenic changes were flanked by unrelated partial genes (Fig. 1). In fewer instances, the changes were associated to chromosomal rearrangements, such as is the case for genes B01–B03 in B19N3. In a few cases, tandem duplicates could be observed in *B. oleracea*, for example genes D06a and D06b, H01a and H01b. A similar situation was observed when *B. rapa* and *A. thaliana* were compared. Clone KBrB063K02 corresponding to chromosome II of *A. thaliana* had two

Table 3 Summary of features identified in seven *B. oleracea* contigs and five *B. rapa* BAC clones

Feature	*B. rapa* BAC clone						
	B. oleracea contigs[a]	KBrH0 93K03	KBrB0 21P11	KBrS0 05I11	KBrB0 80C12	KBrH077A05	Total
Sequence length (kb)	2200	78.3	116	127.7	138.3	113.3	574
Gene models in *B. oleracea* or *B. rapa*	539	14	24	19	21	25	103
Gene density (kb/gene)	6.6	5.6	4.8	6.7	6.6	4.5	5.6
Sequence length in *A. thaliana* (kb)	498	79.9	122.0	103.8	100.8	107.8	514
Gene models in *A. thaliana*	271	28	36	30	25	33	136
Sequence ratio (*Brassica*/*A. thaliana*)	1.70	0.98	0.95	1.23	1.37	1.05	1.12
Number of collinear genes	177	9	19	15	14	16	73
Collinear genes (%)	144	32	52	50	56	49	52
Gene fragments	10?	2	0	1	1	3	7
Masked bp (%)	18	18	7	16	13	13	13
Predicted transposons	69	14	10	21	11	13	69
TE insert into collinearity region	14	5	7	6	4	1	23
Percentage	25	36	70	29	36	8	33

[a] Data extracted from Town et al. (2006)

homologs inserted from chromosomes V and III, respectively and two from homologs on chromosomes IV and I, respectively (Fig. 1). Homologs corresponding to contiguous genes At1g15690 and At1g15700 were segmentally duplicated in *B. rapa*, a few genes downstream the original location next to a retroelement.

We could compare *B. oleracea* and *B. rapa* for segments corresponding to four sets of BAC clones. The general theme of missing genes in either *B. rapa* or *B. oleracea* could be observed as the main cause for collinearity disruptions. For *B. oleracea* BAC B16J1, there was almost perfect colinearity with *B. rapa*, spanning from the homologs of At1g24030–At1g24140, except for At1g24050 that is missing in *B. oleracea*. Four other homologs next to At1g24050, which includes At1g24060–At1g24080, were missing in both the *Brassica* species. Similar to *A. thaliana*, *B. oleracea* genes A09, A10 and A11, which are homologous to genes of At2g32430, At1g70140 and At1g67020, respectively, were missing in the corresponding segments of *B. rapa* (Fig. 1). Collinearity for the segment corresponding to B21F5 was almost identical in both species, except for the presence of gene C03, corresponding to homolog (At2g13865) and a partial gene in *B. oleracea* and absent in both *B. rapa* and A. thaliana (Fig. 1). For BAC clones B67C16, it has higher collinearity with its corresponding sequence in *A. thaliana* than with that of *B. rapa*. Genes G03 and G04 were absent in *B. rapa*, whereas the homolog corresponding to gene At2g25460 was missing in *B. oleracea*, and the segment spanned by genes G05–G08 in *B. oleracea* was absent in *B. rapa*. At least four other segments present in *A. thaliana* were absent in *B. rapa*, At2g25540–At2g25570, At2g25580 and At2g25590, At2g25605 and At2g25610, and At2g25625–At2g25670 (Fig. 1). No *B. oleracea* sequence was available to tell whether these segments were also absent in the corresponding segment for this species. The number of absent genes in *B. rapa* for the segment corresponding to *B. oleracea* BAC clone was very extensive. *B. oleracea* genes H01, H03, H05 H09 and H10 were absent in *B. rapa* (Fig. 1). Additionally, by comparing the corresponding segments between *B. rapa* and *A. thaliana*, two blocks of genes were missing in the former species, At1g16260–At1g16320 and At1g16340–At1g16370 (Fig. 1).

Transposable elements

We detected 95 TEs in the eight *B. oleracea* BACs corresponding to a masked base percentage of 13%, whereas in *B. rapa* these numbers were much lower, 30 and 6.4%, respectively (Table 4). The TE density in the *B. oleracea* BACs was 0.13 (95/746 kb), whereas in the corresponding segments of *B. rapa* was 0.06 (30/495 kb).

Also, we detected 243 TEs in the five *B. oleracea* contigs corresponding to a masked base percentage of 18%, whereas in *B. rapa* these numbers were much lower, 34 and 13%, respectively (Table 5). The TE density in the *B. oleracea* contigs was 0.16 (243/1518 kb), whereas in the corresponding segments of *B. rapa* it was 0.12 (69/574 kb).

In the *B. oleracea* BACs the percentage of retroelements (class 1 TEs) is 61%, and of DNA transposons (class 2 TEs) is 39%. The opposite is true for *B. rapa* where the percentage of retroelments is 34 and 66% for DNA

transposons. For the *B. oleracea* contigs, we found similar frequency of these types of elements, class 1 TEs is 65% and class 2 TEs is 35%. However, for the corresponding *B. rapa* segments of the *B. oleracea* contigs, the percentage for class 1 and 2 TEs were nearly the same (51 and 49%).

In the total *B. oleracea* sequence analyzed, there are 0.7 TEs per gene and 0.15 TE per 1 kb of sequence, and 218 masked bases per 1 kb of sequence. In the *B. rapa* BACs there are 0.5 TEs per gene and 0.09 TEs per 1 kb of sequence, 96 masked bases per 1 kb of sequence. The number and masked base percentage of class 2 TEs is more than the class 1 TEs in both *Brassica* species (Table 6). Sixteen percent of *B. oleracea* sequence, 10% of *B. rapa* sequence and 4% of *A. thaliana* sequence corresponds to TE.

The LTR elements were the main type in the retroelements in both *Brassica* species. The En-spm type is predominant in the DNA transposons of *B. oleracea* and the hAT type is predominant in *B. rapa*.

In the eight *B. oleracea* BACs, 17 transposable elements, 13 DNA transposons and four retroelements were inserted into genes. Only one retroelement inserted into a *B. rapa* gene, which is the ortholog for At1g16170 (Fig. 1).

The insertion of TEs was not frequently associated to chromosomal segments displaying breaks in synteny among species. Thirty one percent of TEs in eight *B. oleracea* BACs and 49% in the contigs were inserted into regions maintaining collinearity with *A. thaliana*, whereas 23% (7/30) of TEs were inserted into regions maintaining collinearity in four *B. rapa* clones with *A. thaliana* (Table 2; Fig. 1).

Little conservation of transposable elements insertions was observed among the three species. Only one SINE type TE with same sequence was found in the corresponding location in *B. oleracea* contig B and *B. rapa* BAC KBRH093K03. This TE is of the AtSB6 type and has 68 bp.

Discussion

Differential gene content for corresponding segments in the three species

Most of the comparative genomics work done to date among *Brassica* species with reference to *A. thaliana*, are based on physical and genetic mapping procedures (Rana et al. 2004; Parkin et al. 2005; Park et al. 2005). Comparative sequencing of specific chromosomal regions provides useful new information on gene density, synteny and conservation of gene collinearity along these segments (Gao et al. 2004, 2005, 2006; Yang et al. 2006; Town et al. 2006). In general, gene density for the chromosomal segments studied was highest for *A. thaliana*, followed by *B. rapa* and *B. oleracea*. The limited sequencing data of *B. rapa* available for this study did not allow us to ascertain orthology with *B. oleracea* BAC clones B47M9, B67C16, B77C13, B59J16 and B16J1. However, the conclusion that can be reached from our survey is that for the chromosomal segments studied, gene density is lower in the *Brassica* species than in *A. thaliana*. In terms of genome expansion, *Brassica*/*Arabidopsis* sequence length ratios for the *B. oleracea* sequence analyzed is 1.7, ranging from 0.86 to 3.2 (Tables 2, 3). However, gene density is not uniform across the genome where regions of higher and lower gene density might co-exist, such as is the case for the region covered by BAC clone B21H13. A similar situation is observed for *B. rapa*, although gene density in this case is higher than in *B. oleracea* (Tables 2, 3). Lower density in the *Brassica* species is associated to larger introns and spacers and to extensive gene rearrangement resulting in the absence of genes in otherwise collinear chromosomal segments with *A. thaliana*. The fact that approximately 50% of the genes absent in the compared segments which can be accounted for the *Brassica* data bases (which are incomplete), indicates that most of these genes have not been lost and are somewhere else in the genomes of *B. rapa* and *B. oleracea*. Due to these rearrangements, the breaks in synteny between *Brassica* and *A. thaliana* can be quite extensive, depending on the chromosomal segment compared. This is also true between the two *Brassica* species, which evolutionarily are considered to be in the same lineage (Warwick and Black 1991). This is in agreement with the results of Parkin et al. (2005) who estimated 74 gross rearrangements taking place between the A and C genome chromosomes of *B. napus*. Using RFLP markers, they were able to identify 21 conserved regions of *A. thaliana* duplicated and rearranged in the A and C genome chromosomes of *B. napus*. These conserved segments were on an average 9 Mbp in length. Physical mapping studies by Park et al. (2005) comparing specific chromosomal segments for all three species; report also breaks in synteny mostly due to gene absence. The earlier report of Kowalski et al. (1994) had already suggested that conservation of synteny and gene content between *A. thaliana* and *Brassica* was limited to specific segments or genomic islands. However, as additional comparative sequencing data are accumulated, it is evident that these conserved islands are small. Furthermore, in spite of their phylogenetic close proximity, the genomes of *B. oleracea* and *B. rapa* have also undergone extensive structural changes resulting in segmental conservation of collinearity. Most of the changes observed are species specific, including gene duplications. Thus, these

Fig. 2 The comparison map of five *B. oleracea* contigs and five *Brassica rapa* BACs clones with *A. thaliana* ▶

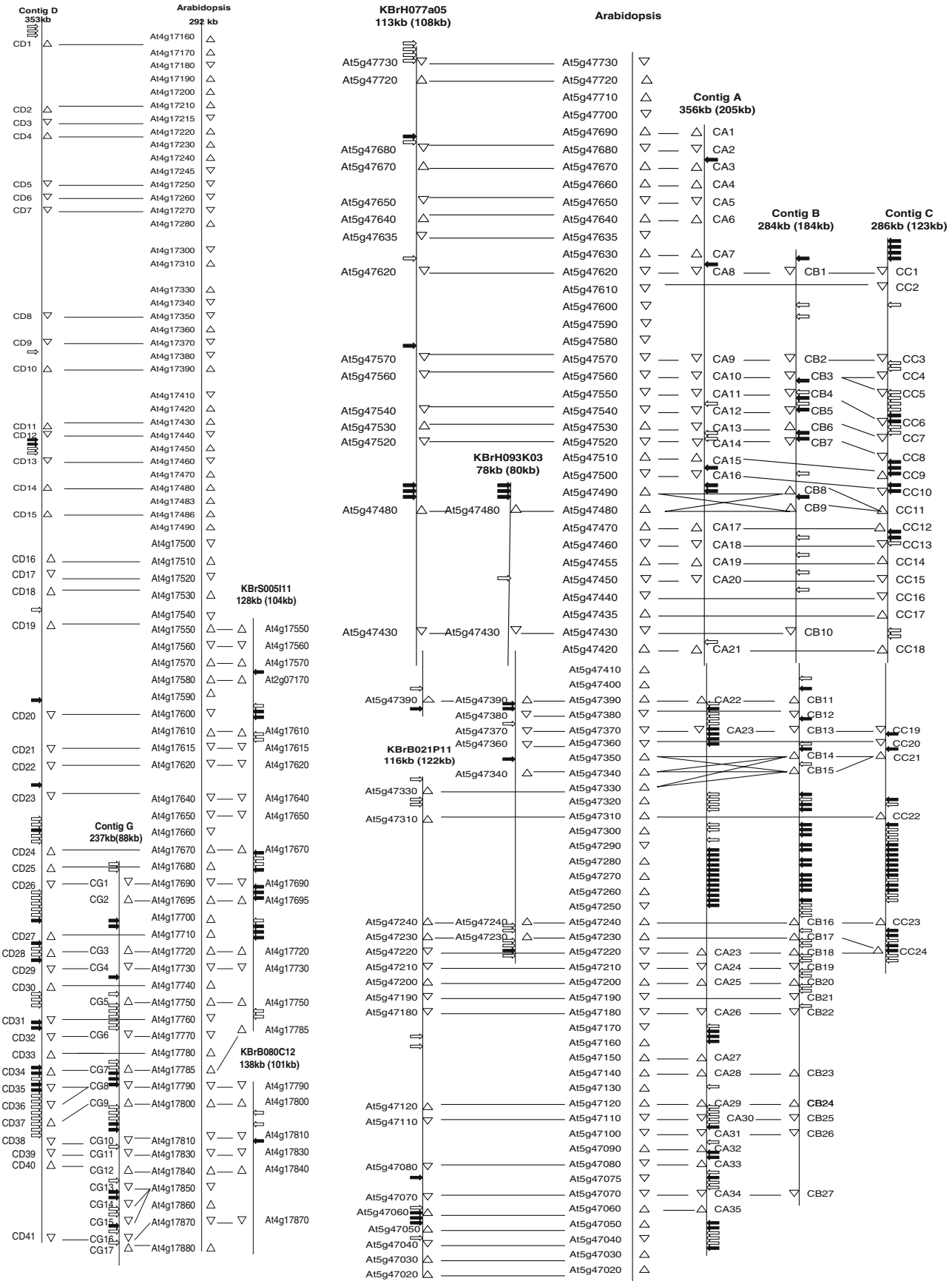

Contig D
353kb
Arabidopsis
292 kb
CD1
CD2
CD3
CD4
CD5
CD6
CD7
CD8
CD9
CD10
CD11
CD12
CD13
CD14
CD15
CD16
CD17
CD18
CD19
CD20
CD21
CD22
CD23
CD24
CD25
CD26
CD27
CD28
CD29
CD30
CD31
CD32
CD33
CD34
CD35
CD36
CD37
CD38
CD39
CD40
CD41
Contig G
237kb(88kb)
CG1
CG2
CG3
CG4
CG5
CG6
CG7
CG8
CG9
CG10
CG11
CG12
CG13
CG14
CG15
CG16
CG17
KBrS005I11
128kb (104kb)
KBrB080C12
138kb (101kb)
KBrH077a05
113kb (108kb)
KBrH093K03
78kb (80kb)
KBrB021P11
116kb (122kb)
Arabidopsis
Contig A
356kb (205kb)
Contig B
284kb (184kb)
Contig C
286kb (123kb)

Table 4 Transposable elements of eight *B. oleracea* and four *B. rapa* BAC clones

Feature	*B. oleracea* BAC clone									*B. rapa* BAC clone				
	B16J1	B19N3	B21F5	B21H13	B47M9	B59J16	B67C16	B77C13	Total	KBrH0 15M19	KBR0 77F22	KBrB0 63K02	KBrH1 070K21	Total
Bases masked (bp)	12861	22849	11597	8261	4368	12937	3377	21605	97855	9935	3358	6765	11232	31290
Percentage	12.92	23.62	14.09	8.28	8.6	16.94	7.94	19.2	13	7.2	3.1	4.9	10.2	6.4
Retroelements	5/6963	10/11092	2/284	2/3458	1/690	7/8949	1/670	6/8913	34/41019	4/2148	2/757	3/260	1/316	10/3481
SINEs	1/63	2/138	2/284	1/111	0	1/155	0	0	7/751	1/63	0	2/190	0	3/253
Penelope	0	0	0	0	0	0	0	0		0	0	0	0	0
LINEs	0	2/627	0	1/3347	1/690	0	1/670	1/99	6/5433	0	2/757	1/70	1/316	4/1145
LTR elements	4/6900	6/10327	0	0	0	6/8794	0	5/8814	21/34835	4/2148	0	0	0	4/2148
DNA transposons	9/2605	12/7008	13/8844	7/1013	6/2056	4/955	5/1014	5/1728	61/25223	4/647	2/109	4/549	10/4983	20/6288
hobo-Activator	1/103	1/49	4/1373	5/741	0	0	0	0	11/2266	2/346	0	1/196	5/2108	8/2650
Tc1-IS630-Pogo	0	2/704	1/188	2/272	0	3/878	0	0	8/2042	2/178	0	1/191	0	3/369
En-Spm	1/59	2/2392	4/4805	0	3/686	0	1/158	1/191	12/8291	0	0	0	2/2152	2/2152
MuDR-IS905	1/671	5/3562	0	0	2/375	0	1/476	1/42	10/5126	0	2/109	0	0	2/109
PiggyBac	0	0	0	0	0	0	0	0		0	0	0	0	0
Tourist/harbinger	1/997	2/2745	3/2923	0	0	1/118	2/241	1/1226	10/8250	0	0	0	0	0
Other	0	0	0	0	0	0	0	0		0	0	0	0	0
Total	14	22	15	9	7	11	6	11	95	8	4	7	11	30

Table 5 Transposable elements of five *B. oleracea* contigs and five *B. rapa* BAC clones

Feature	*B. oleracea* contigs		*B.rapa* BAC clone				
	Total	KBrH0 93K03	KBrB0 21P11	KBrS0 05I11	KBrB0 80C12	KBrH077A05	Total
Bases masked (bp)	397813	14027	7251	20673	18391	14633	74975
Pecentage	18	18	7	16	13	12.9	13
Retroelements	114/147617	6/6029	4/848	14/14646	4/1004	6/8919	34/31446
SINEs	6/606	1/71	1/135	2/321	2/282	2/133	6/942
Penelope	0/0	0/0	0/0	0/0	0/0	0	0/0
LINEs	29/33842	1/806	3/713	2/737	2/722	1/3933	7/6911
LTR elements	79/113169	4/5152	0/0	10/13588	0/0	3/4853	17/23593
DNA transposons	129/79245	8/2160	6/2153	7/10151	7/7505	7/1734	35/23703
hobo-Activator	37/16174	3/1214	0/0	2/307	1/461	5/1251	10/3233
Tc1-IS630-Pogo	10/2332	3/635	5/924	4/725	2/414	2/1519	14/4217
En-Spm	20/43888	0/0	0/0	0/0	2/6185	0/0	0/6185
MuDR-IS905	17/7254	0/0	0/0	0/0	0/0	0/0	0/0
PiggyBac	1/1310	0/0	0/0	0/0	0/0	0/0	0/0
Tourist/harbinger	4/5666	0/0	1/1205	0/0	0/0	0/0	1/1205
Total	243	14	10	21	11	13	34/31446

Table 6 Distribution of transposable elements in *Brassica*

Item	Total TE (num/bp)		Retroelments		DNA transposons	
	B. oleracea	*B. rapa*	*B. oleracea*	*B. rapa*	*B. oleracea*	*B. rapa*
Total TE content	338/495668	99/91632	148/188636	44/26008	190/104468	55/28257
TE per *Brassica* gene	0.7/1043	0.5/569	0.3/397	0.2/162	0.6/219	0.5/175
TE per kb	0.15/218	0.09/96	0.07/83	0.04/27	0.08/46	0.04/29

changes have taken place after the separation of the *oleracea* and *rapa* lineages.

Differential TE content in *B. oleracea* and *B. rapa*

For the segments analyzed, we found a lower frequency of TEs in *B. rapa* than *B. oleracea*, which is in agreement with the smaller genome size of the former species. We estimated that for these segments, approximately 16% of *B. oleracea* sequence and 10% of *B. rapa* sequence consist of TEs, which is not far from a global TE estimate of 20% for *B. oleracea* by Zhang and Wessler (2004) and Katari et al. (2005). Of these, approximately 14% correspond to retroelements and 6% to DNA transposons. For *B. rapa*, based on 60 Mb BAC end sequences, 12.3% of the sequences consist of TE sequences, of which 84% are retroelements and 11.4% are DNA transposons (Lim et al. 2006). Our estimate is also in agreement with this report. Considering that the TE content in *A. thaliana* is only 4% (Zhang and Wessler 2004) the accumulation of these elements has taken place after the separation of the *Arabidopsis* and *Brassica* lineages. Alix and Heslop-Harrison (2004) have analysed the diversity of retroelements in diploid and allotetraploid *Brassica* species, where there is a distinct clustering of Copia-like retroelements in C genome much more than the A and B genome. This result is in agreement with our observation, 83 bp retroelements in 1 kb *B. oleracea* sequence is more than the 27 bp in *B. rapa*.

When comparing TE insertions between *B. oleracea* and *B. rapa*, we found little conservation of TE elements. Only one insertion was shared between the two species. Therefore, it is evident that each species have followed their own path of TE acquisition, where the rate of accumulation of these elements has been higher in *B. oleracea* than in *B. rapa*.

Based on previous reports and our work, TE elements are ubiquitous in *Brassica* species and have an important role in genome evolution. Most likely current estimates for these elements will increase as progress is made on sequence annotation. For example, Lim et al. (2007) after analyzing close to 88,000 BAC clones have found that retrotransposons are major components of centromeres and peri-centromeric regions in most *Brassica* species. One can estimate the contribution of these elements as part of the *Brassica* genomes as follows: *B. oleracea* has a DNA content of 696 Mbp, of which 20% (139 Mbp) are TEs.

Assuming an even distribution of DNA in all nine chromosomes, each has then approximately 77 Mbp of DNA. Thus, the increase in DNA by TEs in *B. oleracea* is equivalent to adding two chromosomes to its genome. If the ancestral *Brassica* lineage had $2n = 4x = 16$ (Henry et al. 2006), we will have to add another eight chromosomes to produce an hexaploid (Lysak et al. 2005; Parkin et al. 2005; Ziolkowski et al. 2006; Yang et al. 2006), which presumable would then have $2n = 6x = 24$. Therefore, a simpler explanation is to assume that the *Brassica* lineage diverged from the tetraploid ancestral lineage proposed by Henry et al. (2006) after insertion of TE elements, segmental duplications and chromosomal rearrangements resulting from hybridization events. This is a more likely scenario to explain the observed regional triplication than invoking another round of polyploidization, followed by massive chromosome loss to return to the existing chromosome numbers of $2n = 16$–20, characteristic of the monogenomic *Brassica* species.

Acknowledgments We are indebted to Mr. Vincent D'Antonio and to Mrs. Fengliang Huang for technical assistance and to Dr. Roger Chetelat for his useful comments on the manuscript. The project was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number to CFQ # 2005-35301-15886 "Cloning and characterization of the major genes involved in the aliphatic glucosinolate pathway of *Brassica* crops".

References

Alix K, Heslop-Harrison JS (2004) The diversity of retroelements in diploid and allotetraploid *Brassica* species. Plant Mol Biol 54:895–909

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. Genome Res 15:487–495

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell 12:1093–1101

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Flicek P, Keibler E, Hu P, Korf I, Brent MR (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. Genome Res 13:46–54

Gao M, Li G, Yang B, McCombie WR, Quiros CF (2004) Comparative analysis of a *Brassica* BAC clone containing several major aliphatic glucosinolate genes with its corresponding *Arabidopsis* sequence. Genome 47:666–679

Gao M, Li G, McCombie WR, Quiros CF (2005) Comparative analysis of a transposon-rich *Brassica oleracea* BAC clone with its corresponding sequence in *A. thaliana*. Theor Appl Genet 111:949–955

Gao M, Li G, Potter D, McCombie WR, Quiros CF (2006) Comparative analysis of methylthioalkylmalate synthase (MAM) gene family and flanking DNA sequences in *Brassica oleracea* and *A. thaliana*. Plant Cell Rep 25:592–598

Henry Y, Bedhomme M, Blanc G (2006) History, protohistory and prehistory of the *Arabidopsis thaliana* chromosome complement. Trends Plant Sci 11:267–273

Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. Mol Cells 22:300–307

Johnston JS, Pepper AE, Hall AE, Chen ZF, Hodnett G, Drabek J, Lopez R, Price HJ (2005) Evolution of genome size in *Brassicaceae*. Annals Bot 95:229–235

Katari MS, Balija V, Eilson RK, Martienssen RA, McCombie WR (2005) Comparing low coverage random shotgun sequence data from *Brassica oleracea* and *Oryza stiva* genome sequence for their ability to add to the annotation of *Arabidopsis thaliana*. Genome Res 15:496–504

Kowalski SP, Lan TH, Feldmann KA, Paterson AH (1994) Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. Genetics 138:499–510

Lagercrantz U (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics 150:1217–1228

Lim YP, Plaha P, Choi SR, Uhm T, Hong CP, Bang JW, Hur YK (2006) Toward unraveling the structure of *Brassica rapa* genome. Physiol Plant 126(4):585–591

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, Kim J, Choi BS, Lim MH, Jin M, Kim HI, de Jong H, Bancroft I, Lim Y, Park BS (2007) Characterization of centromere and pericentromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J 49:173–183

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe *Brassiceae*. Genome Res 15:516–525

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J 23:233–243

Park JY, Koo DH, Hong CP, Lee SJ, Jeon JW, Lee SH, Yun PY, Park BS, Kim HR, Bang JW, Plaha P, Bancroft I, Lim YP (2005) Physical mapping and microsynteny of *Brassica rapa* ssp. *pekinensis* genome corresponding to a 222 kb gene-rich region of *Arabidopsis* chromosome 4 and partially duplicated on chromosome 5. Mol Genet Genomics 274:579–588

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics 171:765–781

Rana D, van den Boogaart T, O'Neill CM, Hynes L, Bent E (2004) Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives. Plant J 40:725–733

Schmidt R, Acarkan A, Boivin K, Clarenz O, Rossberg M (2003) The sequence of the *Arabidopsis* genome as a tool for comparative structural genomics in *Brassicaceae*. In: Nagata T, Tabata S

(eds) Biotechnology in agricultural and forestry (BAF), vol 52. Springer, Heidelberg, pp 19–38

Schranz EM, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the *Brassicaceae*: building blocks of crucifer genomes. Trends Plant Sci 11:1360–1385

Sikka SM (1940) Cytogenetics of *Brassica* hybrids and species. J Genet 40:441–509

Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ, Vigouroux M, Trick M, Bancroft I (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18:1348–1359

Warwick SI, Black LD (1991) Molecular systematics of *Brassica* and allied genera (subtribe *Brassicinae*, *Brassiceae*)—chloroplast genome and cytodeme congruence. Theor Appl Genet 82:81–92

Wroblewski T, Coulibaly S, Sadowski J, Quiros CF (2000) Variation and phylogenetic utility of the *Arabidopsis thaliana* Rps2 homolog in various species of the tribe *Brassiceae*. Mol Phylogenet Evol 16:440–448

Yang YW, Lai KN, Tai PY, Li WH (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and the other angiosperm lineages. J Mol Evol 48:597–604

Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim HI, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS (2006) Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. Plant Cell 18:1339–1347

Zhang X, Wessler S (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. PNAS 101:5585–5594

Ziolkowski PA, Kaczmarek M, Babula D, Sadowski J (2006) Genome evolution in *Arabidopsis/Brassica*: conservation and divergence of ancient rearranged segments and their breakpoints. Plant J 47:63–74